DPC Briefing Day: Email how hard can it be? (2)

Woburn House, 24 January 2018,
http://www.dpconline.org/events/briefing-day/email-2018

Twitter and other online notes

Twitter #DPCEmail2 - within this David Underdown tried to keep his tweets together in two main threads, first thread covers up to the end of Tim Gollin's talk (the threading didn't always quite work, so if you see a tweet has a couple of replies, you may need to click on that tweet indvidually to see the full thread from that point) and then second thread covers remainder of the day.

James Baker's notes

1030 – Welcome and Introductions (William Kilbride)

1035 – Introductory talk (Chris Prom, University of Illinois Urbana Champaign and Kate Murray, Library of Congress)

Brief reminder of the task/purpose of the taskforce
References and bibliography http://www.emailarchivestaskforce.org/bibliography/ (prepared by research assistant from the Carnegie-Mellon funding)

- Need to build user community, and show growing research interest
- Challenge of sensitivity
- Processing at scale
- Using AI/Machine learning for classification and identification of sensitivities, determining what is truly a record
- Reference research about what users (ie researchers) want out of email

- Comments from public review
    - Report should serve as a call to arms regarding importance of access to email archives and the exploration of technical requirements
    - Upskilling workforce
    - Sustainability plans for open source tools

1. Section 1: Untapped Potential of Email Archives
    a. Email matters: even small institutions already receiving, but even large institutions can have trouble in getting buy-in for preservation, accounts often auto-deleted when staff leave. Often perceived as a legal risk.
    b. Changes in archival processes:
        i. embrace email as complex research data, embrace new tech such as NLP/ML: we can do things with email that would never be possible for paper.
        ii. Encourage those who create the content to take active role, even for their personal papers. Very hard when we transfer from one institution to another
        iii. Build toward tool interoperability and community integration.
    c. Some contrast between email as an organisational record v personal record
        i. Former may have some sort of record management structure, and are documenting business actions, decisions etc, latter perhaps less structured, covering a wider range of activities.
2. Section 2?
3. Section 3:: Email as a Documentary Technology
    a. What is email, actually?
        i. A verb
        ii. An individual message
        iii. The underlying transmission mechanism etc
        iv. Attachments, links etc
        v. Where are the boundaries?
        vi. Email multiplies, who's copied in, what's attached at any time, often have multiple versions (incorporating earlier parts of chain) saved in corporate RM solutions
4. Section 4: Current Services and Trends

    a. Evolving email ecosystem: abuse prevention (ie cutting down on spam), adding authentication (and then how, and need we, preserve that context too?)
    b. Repository challenges:
        i. Capture
            1. Direct export
            2. Web service exports
            3. Client-based exports
            4. D
        ii. Attachments
5. Section 5: workflows and potential solutions
    a. Preservation approaches
        i. Begins with account and format analysis
            1. Sufficiently standardised that tools do exist for migration from most storage formats
        ii. Bit-level: possibly appropriate for embargoed collections
        iii. Format migration
            1. Many tools are format dependent
            2. Always risk/reward
            3. Some formats play better together than others
            4. Open, non-proprietary formats perceived better than closed, proprietary formats
        iv. Emulation
6. Section 6 Tools within the Cultural Heritage Domain
    a. Key to interop, scalability and acccess
        i. Epadd, various imports, MBOX, PST; entity extraction/NLP tools, discovery and delivery environment
        ii. TOMES, cross-platform PST to EAXS XML parser, capstone processing, NLP named entities particular to state/local government, training materials http://www.ncdnr.gov/resources/records-management/tomes
        iii. Harvard Electronic Archiving System - EAS, http://library.harvard.edu/preservation/email-archiving more end-to-end, will identify attachements and migrate etc maintaining reference from email
        iv. Proprietary tools

1120 – Q&A

Possible follow on, making some an initial range of collections available for researchers to start to get to grips with. William mentions Archaeology Data Service, can't think of a single collection there which includes email, in contrast with existing collections of archaeologists' private papers. Natalie Harrower picks up on the idea of treating email as research data, rather than just as analogous to correspondence, and what does that mean in practice? Chris Prom: Functions of RDM, archives etc within universities often somewhat separate currently. Does treating email as data make it less scary? Or at least move it to a better resourced area? Tim Gollins, we must be careful not to "damage" the material we have received by "destructive" processing before we ingest, need to be able to go back to "original" form and apply different processes as they become available. CP comments that this is rather buried as an assumption behind some of the workflows etc. [Justine Mann? or Marta Fernandez Campa? - UEA Archives] Stressing need for more user friendly tools for personal use (in context of literary archives)

1130 – Using email archives in Research (James Baker, University of Sussex)

James says that his interest was in part sparked by reading Born-digital archives at the Wellcome Library: appraisal and sensitivity review of two hard drives Victoria Sloyan Pages 20-36 | Published online: 25 Apr 2016 (open access) and also Carolyn Steedman's book, *Dust*

Digital archives are not unique: infinitely reproducible.

Some researchers already looking at the contemporary social history, perhaps not directly grappling with email yet.

Also literary researchers eg Matt Kirschenbaum, Doug Reside, Thorsten Rise. Again coming from "literary" side, as personal papers, not using the features of email specifically.

Ethical considerations: researchers not always explicitly trained in this field, but does tend to why researchers tend to focus on slightly older material where these issues are less to the fore.

From this year undergrads largely born 2000 on, the 1990s are now truly history?? So digital material more likely to be investigated?

Out of this came the born-digital access workshop with Collections Information team at Wellcome (in Simon Demissie, Alexandra Everleigh, Victoria Sloyan), 24 November 2017. 4 case studies, approx 30 minutes on each. A “Original environment”, using original Windows laptop, browsing hard drive etc. B “viewer”, treating the digital files more like digitised objects. C “levels of curation”, renamed files etc, was there enough to work with. D adding more on physical media and how that influenced access (printed emails?), how well are paper, physical media, digital files and inline catalogue connected? (None actually included emails).

Themes from responses on the day, levels of description, using a laptop as access point, getting researcher “from desk to archive”.

Levels of description, original filenames/directory structures perceived as important data in own right. Lack of curation hard to deal with. Wanted multiple categorisations. Wanted high level view of collection attributes desired. Just want access. Some of these seem a little contradictory. Also desire to know high-level view of file size, file formats etc

Laptop as access point. Modern OS as useful research tool. Archival research doesn’t need dust to feel right. Uncanny illusion of familiarity. But if not actual original environment does this truly reflect original use/experience

Getting from desk to archive. How would researcher know they wanted to use the archive. How would I know what to request. Item level descriptions too much if they aren’t detailed. Remote search> Non-disclosive analysis offsite. Expectation raised when archives not paper.

Tension between participants who valued original order etc (filenames etc) v those who wanted curation. Few questions about accessing digital records in new ways, still tending to use in same way as paper. Discussions among archivists haven’t really made it out into research community yet.

1200 – The reconstruction of narrative in E-Discovery investigation (Simon Attfield, Middlesex University and Larry Chapin, Attorney)

Larry Chapin is a consulting attorney, among other cases he was involved in eDiscovery on some of the US investigations into LIBOR (London inter-bank offer rate) rate fixing. This investigation took over three years.

He started with a brief definition of eDiscovery:

> *A process in which electronic data is sought, located, secured and searched with the intent of using it as evidence in civil or criminal proceedings, or as part of an inspection ordered by a court or sanctioned by government.*

Chapin believes that lawyers need to think in terms of narrative to recover the real meaning of the documents returned via eDiscovery. When a case begins lawyers don't really know that much about what happened, at least in terms of the real detail.

For LIBOR, the story of the investigation is kicked off by a 2008 *Wall Street Journal* article calling the rate setting process into question.

Other cases relating to LIBOR had led to Barclays having to make a multi-million pound settlement with regulators. In that case the evidence showed that the fixing was top-down: pressure from senior execs led to those within the bank responsible for reporting the bank's rate into the LIBOR system declaring lower rates than they were actually obtaining (which made the bank appear healthier).

In the cases with which Chapin was particularly involved they were expecting to find similar downward pressure, but the emails told a different story. The pressure was much more bottom up, or between peers. Rather than being on an institutional level, the activity was really about improving the profit and loss account of individual traders (and hence getting them bigger bonuses!).

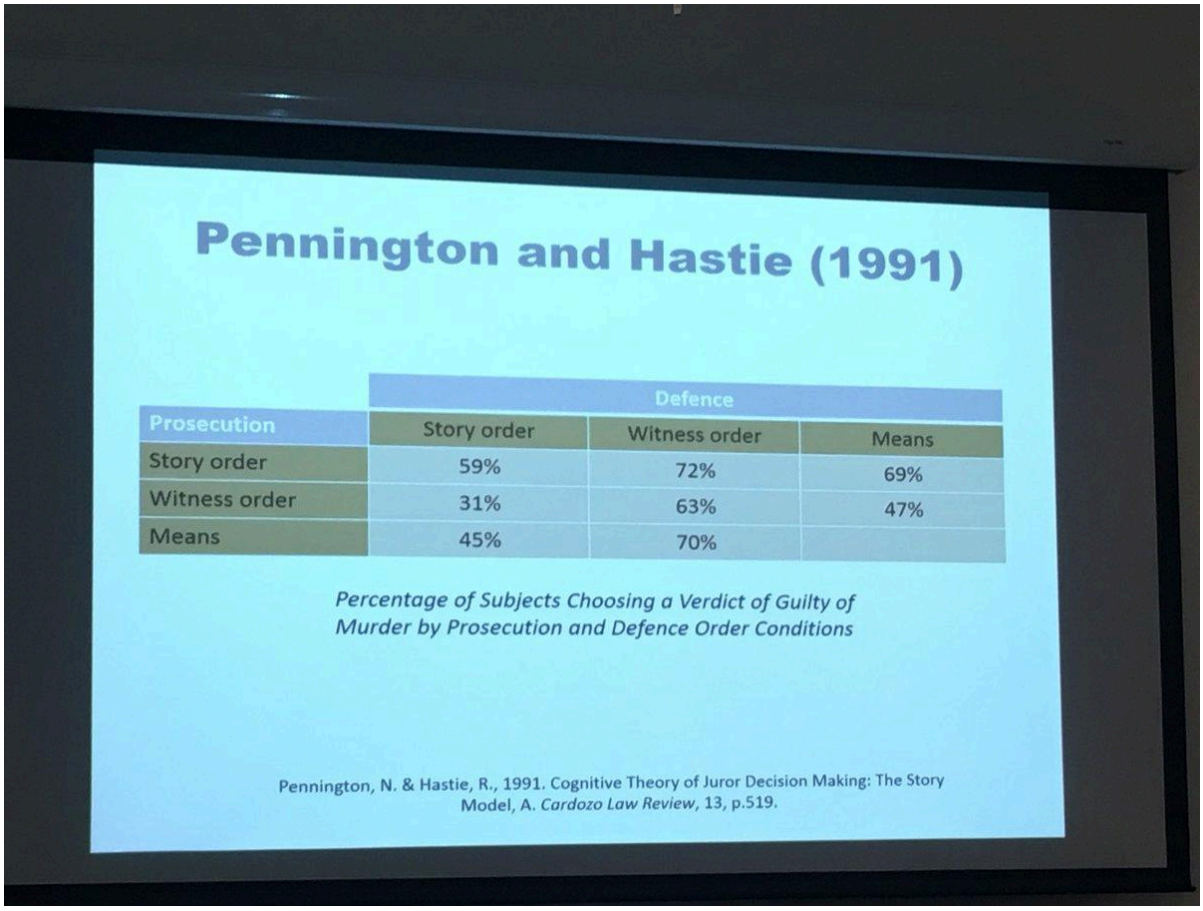
Emails [and perhaps now other messaging tools] a motherlode in reconstructing events. Email is quite a democratic (or egalitarian tool), generally it is now available to all staff within an organisation. As a result it's often treated quite informally and messages can be quite jargon filled. They found that some of the most memorable and revealing emails were actually pretty banal in content "thanks I owe you, gonna open a bottle of champagne for you". Emails also revealed unexpected relationships between people, and this helped shape direction of investigation (eg extramarital affairs changing investigators' thoughts on who was colluding). Another confusion ws fairly widespread use of nicknames, or people actually being known by a

very different name to how you'd find them listed in a formal corporate directory eg a William being known as Zac, only by establishing these variances in names was it possible to work out who was actually being referred to when mentioned in others' emails.

One issue is that it's been established that people (investigators or subsequently jurors) can only really hold onto detailed knowledge of around 5-9 documents at a time. So the way you are led from one document to another can have an impact as to how you understand the relationship between various documents.

Simon Attfield then described related academic research. He starts with a very brief description of predictive coding and how it's used. In essence, lawyers present the computer system with some documents know to be related to the case, the system then searches through the entire document collection and returns "similar" documents. The lawyer then reviews what has been returned and decides which were actually relevant, and the process then iterates, with the computer system learning from what the lawyers marks as being truly relevant at each iteration.

Mention of interesting study by Pennington and Hastie 1991 looking at how different presentation of cases by defence/prosecution [in summing up?] affects verdicts. In general explaining by referring back to evidence in a narrative order (rather than in the order witnesses were examined) had better results.



Similarly the very way investigators record their investigation can affect how the investigation proceeds. Attfield/Middlesex University have developed method for scoring "structuredness" of approach. Tested with various people on defined corpora of documents. Found high correlation between highly structured approaches (building narrative) and precision and recall of relevant documents. Also an inverse correlation with the time taken to perform document triage for relevance (ie less time taken when the overall process well structured, so investigators able to make better decisions more quickly).

1230 – Email as corporate record (James Lappin)

[Similar to talk given by James at TNA last year]
General expectation within UK government that emails will be moved out of inbox into EDRMS system, as opposed to CAPSTONE approach now being taken in US where whole email accounts will be preserved. Assertion taht only perhaps 1 in 20 to 1 in 200 emails preserved.

Can often be coupled (TNA Guidelines revised 2016) with auto-deletion of emails after a defined period. In theory this should lead to organisations improving tools etc to ensure a good record of their activities is being kept, but in practice… Results in big issues around ongoing government accountability and highlighted in Sir Alex Allan's review of government record keeping. This report did recommend consideration of CAPSTONE type approach, but Cabinet Office response asserts that this is not practical within UK (European) legislative framework due to data protection considerations. James considers that there are circumstances in which this could be overcome, where:

- The role played by a given civil servant is of historic interest [eg Treasury chief economist]
- They have been given expectation that their account will be permanently preserved
- Given chance to remove/flag personal correspondence
- Access to any personal correspondence prevented unless overriding legal need

Interesting point that part of issue is that email is often less considered, a civil servant now will probably send 10s of emails per day, often a quick off-the-cuff reply, [with the more informal language noted by Chapin in previous session] rather than a handful of carefully crafted memos.

Equally the CAPSTONE approach can also be hamstrung if we cannot be sure that personal material has been removed. This can be perceived as making keeping an email account a liability and risk rather than a valuable asset.

Describes email as leading to the first time where correspondence cannot routinely be passed on to successor in a role. [Tim Gollins points out instances where role-based accounts exist rather than personal accounts eg in MOD, common for email accounts to be defined as eg Chief of the Defence Staff rather than in name of current holder of that position]

1300 – Lunch

1400 – Jason R. Baron, Drinker Biddle LLP (by video conference)

Giving an intro to his four propositions for email. Context, NARA already holds millions of emails, but the personal data problem means it will be years before they are generally publicly accessible. Describes emails with attachments as lingua franca of government over last 25 years.

- Proposition 1: manual approaches lead to non-compliance

Discusses attempts to apply traditional ideas of scheduling emails as records [very similar to much of what James Lappin said]. Worked in paper when we had secretaries/file clerks etc to actually do the work (and lower volume), but not when everyone has to do the filing themselves and with the amorphous volume of email. US effectively still on print-to-paper until 2011 when Obama issued records management order, followed by detailed instructions for NARA etc in 2012. Describes history of failed approaches, print-to-paper, simple backup of email, even the detailed requirements of DoD 5015.2. All leading to:

- Proposition 2: automated capture

In essence the driver for CAPSTONE. Capture entire accounts of "important" individuals, typically senior figures within an agency, but also those relating to other key roles defined on agency-to-agency basis. All other emails treated as temporary and deleted after 7 years (in absence of any requirement for legal holds etc etc). Anticipates this will capture approx 1.5% of emails for permanent preservation [so is UK situation described by James Lappin actually so bad?].

Alternatively push for development of tools for automatic categorisation so that emails can be captured into relevant categories automatically [regardless of who is sending them]

- Proposition 3: Automated tools for access

Need to deploy automated tools to enable access too, ensuring researchers can find the information they need while avoiding issues of personal data and other sensitivities. Thinks eDiscovery tools and related machine learning approaches have definite application here, refers to TREC [Text Retrieval Conference] legal track

- Proposition 4: the sensitivity review problem

Also need to incorporate these tools into our initial archiving workflows to identify problematic material and maybe prepare redactions. US situation is essentially default of 75 year closure when PII present [Kate Murray refers to NARA FOIA screening guidelines to find further info]. Ends with call for further research on this problem. Also looking at the signal to noise ratio from the CAPSTONE approach and auto categorisation and filtering, importantly review how this all feeds back to access for researchers. Also problem that algorithms can be black boxes. How do we manage to make them trusted/understood? [Algorithmic accountability].

Q&A following on, Lee Hibberd (National Library of Scotland) asks about how we can determine if/when algorithms are good enough? Jason says that we have to realise that perfection is unattainable but that algorithms can already be shown to do a better job than straight manual review. Realises that archivists may find this a difficult message [chimes with desire in TNA digital strategy to be better at expressing and capturing uncertainties within/around our records]

1430 – Panel discussion introduced by Tim Gollins (National Records of Scotland): technology assisted review, fact, fiction or jam tomorrow and why this matters for email preservation

Tim begins with a quick review of what TAR [technically assisted review] is.

- The technical:
    - Uses computers, usually supervised learning, system is shown what “good” documents look like then finds “similar”
- Assisted
    - Not the system making decisions by itself, helps the human, not replaces
- Review
    - This approach means that every document is actually checked [by the computer]

Cautions that sensitivity review is not exactly the same task as legal discovery, albeit related. Refers to work at University of Waterloo by McCormack and Grossman on the emails of Governor of Virginia, Tim Kaine and papers by Graham MCDonald at Glasgow looking at FCO telegrams: particular issues around sensitivity only really arising in context eg someone referred to in early records later becomes highly significant person in home country, content of early messages could then have potential to damage international relations.

1515 – Coffee

1530 – Review of Report Recommendations and Roadmap (Chaired by William Kilbride)

Chris Prom presents the initial recommendations of the task force, looking at two main areas, and in each case some relatively easy things, and then a set of harder/more long-term ambitions.

1. Community development and advocacy (nurturing and fostering)
    a. Lower-barrier activities
        i. Assess institutional readiness
        ii. Training and skills development
        iii. Demystify email archiving for collection donors
        iv. Maintain assessment of email tools in COPTR
        v. Develop [email] format comparison matrix
    b. Higher-impact activities
        i. Sustain the email archiving community
        ii. Specification planning for beginning of lifecycle email tools
        iii. Develop criteria for email authenticity
        iv. Improve standards documentation for MBOX and EML
        v. Improve options for PDF in email archiving workflows
2. Tool support, testing and development
    a. Lower-barrier activities
        i. Test existing tools for data impact and data loss
        ii. Improve format identification, characterisation and validation tools for email formats
    b. Higher-impact activities
        i. Sustaining and integrating existing tools
        ii. Develop email self-archiving tool
        iii. Develop standards for tool interoperability with a reference implementation
        iv. Improved tools for sensitivity review

1630 – Next steps

From those high level recommendations we moved into discussion of next steps, Tim Gollins comments that though the task force has (understandably) avoided discussion of legal/policy issues in general, we need to be aware that specific legal frameworks may drive tool requirements eg generally the US approach to removing PII (personally identifying information) is working in a looser framework around privacy law than exists in Europe where existing data protection legislation soon to be replaced by GDPR (General Data Protection Regulation). Also sees that one of the most important things to develop is test corpora in order to have something to develop against. Justine Mann comments that some of this work may actually help us to pushback on some legal requirements to enable useful archival processing.
William Kilbride comments on the comments throughout the day about email being perceived as liability, can we express its value better, paper personal correspondence already valued for insurance, and also the UK acceptance in lieu scheme where donation of heritage material can be written off against tax due. Others ask if we really want market model in born digital collections?

James Baker comments that test corpora and case studies might also help build donor confidence, they can see what is possible to emails, and how personal material kept safe. Kelly Stewart remarks that the corpora would also be useful for wider digital preservation development such as Archivematica by Artefactual.
Michael Day asks about the wider issues of handling attachments and links within emails to documents held elsewhere, whether that is on public internet, intranets or EDRMS.
David Underdown mentions some characterisation issues with MIME attachments, DROID etc use binary signatures, attachments usually base64 encoded so the signatures won't work.
Chris Prom comments that the tools usually extract and decode back to the original binary representation so hasn't seemed too much of a problem [but could be an issue for us if we wanted to treat eg MBOX as a container that DROID could then scan inside].
Tim Gollins then asks about overlap between email and calendars, the two are often pretty integrated. Chris Prom says not explicitly considered, usually some sort of vendor extension to email, so beyond initial scope.

1645 – Close

Thanks to all, particularly speakers and also Kate Murray (Library of Congress), sadly prevented from attending by the US Government shutdown and Fran Baker (University of Manchester) similarly prevented from attending by jury duty, conspiring to leave William faced with a manel much to his embarrassment.